基于多特征提取和 Stacking 集成学习的金线莲品系分类

谢文涌[1,2]，柴琴琴[1,2※]，甘勇辉[3]，陈舒迪[1,2]，张 勋[4]，王 武[1,2]

（1. 福州大学电气工程与自动化学院，福州 350108；2. 福建省医疗器械和医药技术重点实验室，福州 350108；3.漳州职业技术学院食品工程学院，漳州 363000；4. 福建中医药大学药学院，福州 350122）

摘 要：针对传统中药鉴定、分子鉴定、生物技术鉴定及光谱检测技术的主观性强、耗时、操作复杂等不足，以及金线莲整个叶片形态区分度小、单一分类器鉴别精度不高的问题，该研究提出了基于机器视觉的叶片子区间多特征提取方法和基于多模型融合的 Stacking 集成学习算法实现金线莲的品系分类。试验采集 6 个品系的金线莲叶片图像数据，进行图像预处理后提取叶片子区间内纹理、颜色共 114 个特征，基于这些特征，构建堆叠式两阶段集成学习框架，以逻辑回归、K 最近邻、随机森林和梯度提升决策树（Gradient Boosting Decision Tree，GBDT）作为基分类器，GBDT 作为元分类器进行学习。试验结果表明，Stacking 集成学习模型的整体识别综合评价指标 *F* 值达 93.91%，分类正确率达 94.49%，分别比逻辑回归、K 最近邻、随机森林和 GBDT 这 4 个单一分类模型高出 4.40、11.87、11.01、12.94 个百分点和 5.36、11.34、6.93、12.13 个百分点。因此，该研究能够有效识别金线莲品系，为形状大小相似、形状特征难以利用的植物叶片识别提供参考。

关键词：机器视觉；模型；金线莲；子区间分割；特征提取；Stacking 集成学习；植物叶片

doi：10.11975/j.issn.1002-6819.2020.14.025

中图分类号：TP391 **文献标志码**：A **文章编号**：1002-6819(2020)-14-0203-08

谢文涌，柴琴琴，甘勇辉，等. 基于多特征提取和 Stacking 集成学习的金线莲品系分类[J]. 农业工程学报，2020，36(14)：203-210. doi：10.11975/j.issn.1002-6819.2020.14.025 http://www.tcsae.org

Xie Wenyong, Chai Qinqin, Gan Yonghui, et al. Strains classification of *Anoectochilus roxburghii* using multi-feature extraction and Stacking ensemble learning[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2020, 36(14): 203-210. (in Chinese with English abstract) doi：10.11975/j.issn.1002-6819.2020.14.025 http://www.tcsae.org

0 引 言

金线莲，又名金丝线、金耳环、乌人参、金钱草等，是中国的二级保护植物[1]，被誉为“药中之王”，能够用于预防和治疗高血压，糖尿病，高脂血症，肝炎和肿瘤等疾病[2-3]。由于野生金线莲的自然繁殖率低、生长条件受限制等原因导致数量有限，市面上出售的金线莲大多为人工培育品种。目前，金线莲品系繁多，不同品系的金线莲外观相似但药效差异大，市场上以次充好的现象层出不穷，因此，如何准确有效识别金线莲的品系对保障药方药效、维护消费者利益具有重要意义。

金线莲的品质鉴定通常依赖于化学分析方法，主要包括显微鉴定法、高效液相色谱法、DNA 分子鉴定法和近红外光谱检测技术等，然而这些方法需要在专业人士指导下进行，存在主观性强，过程耗时且精度低，操作过程复杂且费用昂贵等缺陷[4-5]。为了克服以上鉴别方法的缺陷，学者们将机器视觉技术引入到中药材的鉴别中[6-7]，主要集中在通过中药材叶片的识别来判定药材质量，但尚未见有关基于机器视觉技术对金线莲进行品系分类的研究报道。在叶片的特征提取方面，通常使用形状特征[8]、纹理特征[9]和颜色特征[10]来作为叶片的识别特征，为了更加充分表达叶片信息，进一步提高叶片识别精度，特征融合在叶片识别中得到广泛应用[11]。然而上述研究主要是针对不同类别的植物叶片分类进行，对于同种类别不同品系的植物叶片分类研究鲜有报道。

基于提取的特征，构造合适的分类器是通过叶片识别解决品系鉴别问题需研究的另一个重点，常用于叶片识别领域的机器学习方法有 K 最近邻（K Nearest Neighbor，KNN）[12]、支持向量机[13]、逻辑回归（Logistic Regression，LR）[14]、极限学习机[9]等。然而这些识别方法均要求测试数据集与训练数据集的样本概率分布一致，要在众多假设函数构成的空间中确定一个与实际情况最相符合的特征分布函数作为分类器。但是由于金线莲种植基地分布广泛、复杂的生长环境会造成叶片形态异常，如存在如脉纹、叶形异常等情况，实际测试集数据分布存在不确定性，找到一个与实际情况最相符合的分布函数十分困难。因此，单一分类器往往会存在泛化能力不佳的问题。为克服由于样本不确定性带来的分布函数难以精确估计的问题，学者们提出了集成学习方法，通过将多个弱分类器集成为强分类器完成高精度的分类任务。目前，集成学习已成为机器学习的热门研究方向之一[15]，常用的集成学习方法有并行化集成 Bagging[16]，序列化集成的 Boosting[17]和多层分类

收稿日期：2020-04-03 修订日期：2020-06-17

基金项目：国家自然科学基金项目（61773124）；福建省科技厅高校产学合作项目（No.2019Y4009）；福建省食品药品监督管理局金线莲质量标准提升专项（[3500]FJJF[DY]2018008）

作者简介：谢文涌，主要从事图像处理与机器学习研究。

Email：1024396820@qq.com

※通信作者：柴琴琴，副教授，主要从事机器学习与模式识别研究。

Email：qq.chai@fzu.edu.cn

器组合的 Stacking[18]。不同于 Bagging 和 Boosting 集成方式，Stacking 使用高级别的元分类器来综合低级别的基分类器的输出特征以此来增强泛化能力，得到更高的预测精度，降低模型过拟合风险。Stacking 集成学习在教育、医学、社会科学等领域得到广泛应用[19-21]，然而在农业方面的应用甚少[22]。

综上所述，本文提出了基于多模型融合的 Stacking 集成学习算法实现不同品系金线莲叶片的分类。首先针对利用整个叶片形状难以区分不同品系的金线莲的问题，提出了基于子区间分割的颜色、纹理等特征提取和融合方法；其次以提取的特征为基础，在 Stacking 集成框架下设计 LR、KNN、随机森林（Random Forest，RF）[23]和梯度提升决策树（Gradient Boosting Decision Tree，GBDT）[24]作为基学习器，并以 GBDT 作为元学习器输出分类结果；最后，通过算法对比验证所提方法的有效性。

1 材料与方法

1.1 样本图像获取及预处理

1.1.1 图像获取

金线莲含有的多糖和黄酮成分是其药效的有效成分，不同品系的金线莲含量具有显著差异，影响金线莲的售价。目前，市场上主要售卖的金线莲品系分为小圆叶、大圆叶、尖叶、台湾金线莲和杂交金线莲。因此，本研究采集市场上广泛流通的小圆叶、大圆叶、尖叶、台湾金线莲等主要品系以及杂交品系（红霞、一株圆叶）作为研究对象，以此来验证品系分类方法的有效性。

试验所用全部金线莲样本均来源于福建葛园生物科技有限公司，并经福建中医药大学教师专业认证。采集人工套袋培养的 6 个品系金线莲叶片，分别包含台湾金线莲 58 幅、红霞 64 幅、小圆叶 45 幅、尖叶 46 幅、一株圆叶 60 幅、大圆叶 44 幅，总共 317 个样本。为了使采集的样本更具有代表性，将样本平铺于拍摄箱中的白色纸板上，箱内灯光均匀照射在叶片，相机镜头与目标样本的距离为 30 cm 左右。采集设备为尼康单反数码相机，型号为 NIKON D7100，图像分辨率为 6 000×4 000 像素。采集的代表样本图片如图 1 所示。

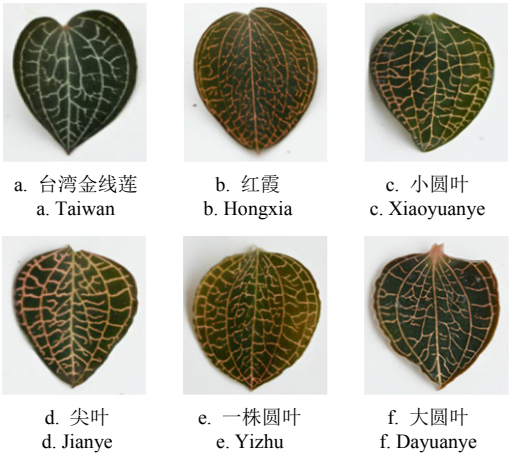


图 1 金线莲不同品系叶片代表图

Fig.1 Representative leaves of different strains of *Anoectochilus roxburghii*

1.1.2 图像预处理与子区间划分

为了避免除了叶片本身其他因素干扰，需对拍摄的原图进行预处理，流程如图 2 所示。首先将图像分辨率缩小至 800×800 像素，先使用均值滤波对彩色图像降噪，其次将图像灰度化，对灰度化后的图像使用高斯滤波去除噪声点；然后用最大类间方差法（OTSU Method）[25]进行自动阈值分割得到二值化图像，再用数学形态学方法中的闭运算消除叶片存在的内部孔洞，开运算去除叶柄；最后，针对不同品系的金线莲形状大小极其相似，叶片形状特征难以利用的问题，提出对叶片进行子区间分割来避免叶片形状特征的影响，在二值图像的基础上绘制原图的叶片轮廓，计算叶片的质心，选择以质心为中心，边长为 150 个像素点的正方形区域作为叶片的子区间，从完整叶片中分割出相同部位和大小的子区间叶片。

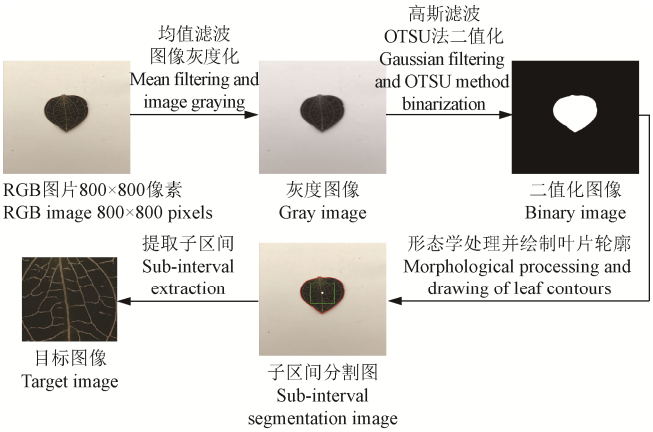


图 2 图像预处理与子区间划分流程图

Fig.2 Flow chart of image preprocessing and sub-interval division

1.2 图像的多特征提取

1.2.1 纹理特征

本文使用基于统计的分析方法局部二进制模式（Local Binary Pattern，LBP）[26]和灰度共生矩阵（Gray Level Co-occurrence Matrix，GLCM）[27]来提取纹理特征。为了更充分地表示纹理特征，使用时域和频域相结合的 Gabor 滤波[28]进行精细比例分析，并对纹理进行多分辨率表示，将这三类特征融合作为金线莲叶片识别的纹理特征。

1）LBP

LBP 作为纹理特征的典型代表，具有计算简单、灰度不变性和旋转不变性等特点。原始的 LBP 计算如式（1）、（2）所示。定义在目标灰度图 3×3 领域内，一个以 (x_c, y_c) 为中心坐标的 3×3 领域的 LBP 值二进制形式会出现 $2^8 = 256$ 种，为了方便计算通常将二进制转化为十进制形式。

$$g(i_p - i_c) = \begin{cases} 1 & i_p \geqslant i_c \\ 0 & i_p < i_c \end{cases} \tag{1}$$

$$\mathrm{LBP}(x_c, y_c) = \sum_{p=0}^{p-1} 2^p g(i_p - i_c) \tag{2}$$

式中 i_c 为中心像素点的灰度值，i_p 为中心点周围等距像素点的灰度值，p 为领域点的个数。

随着采样点数的增加，二进制模式的数量急剧增加，LBP 算子计算量也随之增大。为了解决这个问题，提高统计性，采用等价的 LBP 模式进行降维[29]，这种方法既减少了二进制模式数量又不丢失信息。对于有 8 个领域点的 LBP 算子，二进制模式的数量由 256 种减少至 59 种。

2）GLCM

GLCM 是图像的二阶统计度量，可以反映空间中任意两点灰度值的空间相关性。本文设置图像的灰度级为 16，并将其归一化后分别在 4 个不同的方向（0°,45°,90°,135°）采用对比度 *Con*、熵 *Ent*、能量 *Asm* 和反差分矩阵 ***Idm*** 一共 4 个统计量总共 16 个特征值来反映叶片图像纹理信息。

$$Con=\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}(i-j)^2P(i,j) \tag{3}$$

$$Ent=\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}P(i,j)\log_2(P(i,j)) \tag{4}$$

$$Asm=\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}P(i,j)^2 \tag{5}$$

$$\boldsymbol{Idm}=\sum_{i=0}^{N-1}\sum_{j=0}^{N-1}\frac{P(i,j)}{1+(i-j)^2} \tag{6}$$

式中 $P(i,j)$ 为归一化后的灰度共生矩阵的第 i 行第 j 列的值，N 为灰度级数。

3）Gabor 滤波

Gabor 滤波器能捕获到与图像局部结构信息相对应的不同空间频率、空间位置和方向的特征，被广泛应用于图像中提取纹理特征。在空间域中，二维 Gabor 滤波器是由正弦平面波调制的高斯核函数，具有表示正交方向的实部和虚部。Gabor 滤波器的函数数学表达式如下

$$g(x,y,\lambda,\theta,\psi,\sigma,\gamma)=\exp(-\frac{x'^2+\gamma^2y'^2}{2\sigma^2})\exp(i(2\pi\frac{x'}{\lambda}+\psi)) \tag{7}$$

实数部分为

$$g(x,y,\lambda,\theta,\psi,\sigma,\gamma)=\exp(-\frac{x'^2+\gamma^2y'^2}{2\sigma^2})\cos(2\pi\frac{x'}{\lambda}+\psi) \tag{8}$$

虚数部分为

$$g(x,y,\lambda,\theta,\psi,\sigma,\gamma)=\exp(-\frac{x'^2+\gamma^2y'^2}{2\sigma^2})\sin(2\pi\frac{x'}{\lambda}+\psi) \tag{9}$$

式中 $x'=x\cos\theta+y\sin\theta$，$y'=-x\sin\theta+y\cos\theta$。$\theta$ 表示 Gabor 核函数的方向，x、y 分别表示原图片像素坐标位置，λ 表示波长，cm；ψ 表示相位偏移范围−180°～180°，σ 表示高斯函数的标准差，γ 表示空间宽高比。

在进行特征提取之前，首先要将原图像与 Gabor 函数进行卷积运算来生成与原图像大小一致的目标图像。本文利用 4 个方向，6 个尺度的 Gabor 滤波器生成 24 个卷积模板。将所提取的 Gabor 子带（共 4×6 个）的均值特征组合起来形成一个 24 维的特征向量。

1.2.2　颜色特征

颜色是图像识别中最简单直接的特征，具有良好的鲁棒性。由于颜色低阶矩[30]中含有丰富的颜色分布信息，一阶矩通过计算颜色的均值反映图像的明暗信息，二阶矩用来描述颜色的标准差反映图像的颜色分布范围，三阶矩重点突出颜色的偏移性反映颜色的分布对称性。所以，本文采用低阶矩来提取色彩特征。通过测量整个图像的颜色分布，针对颜色模型 HSV 和 RGB 进行计算。然而 HSV 模型中 V 分量与色彩无关，所以提取颜色模型中的 R、G、B、H、S 共 5 个颜色分量。最后，统计得到 15 个颜色特征。具体的计算公式如下：

$$C_{i1}=\frac{1}{N}\sum_{j=1}^{N}P_{ij} \tag{10}$$

$$C_{i2}=\left(\frac{1}{N}\sum_{j=1}^{N}\left(P_{ij}-C_{i1}\right)^2\right)^{1/2} \tag{11}$$

$$C_{i3}=\left(\frac{1}{N}\sum_{j=1}^{N}\left(P_{ij}-C_{i1}\right)^3\right)^{1/3} \tag{12}$$

式中 C_{i1},C_{i2},C_{i3} 分别表示第 i 个颜色分量的一阶矩、二阶矩、三阶矩，P_{ij} 代表第 i 个颜色分量灰度值为 j 的像素点出现的概率。

2　基于 Stacking 集成学习的金线莲品系识别

2.1　Stacking 集成学习基本原理

Stacking 集成学习框架由两级分类器构成，低级别的分类器称为基学习器，高级别的分类器称为元学习器。训练过程如下：假定原始的一组数据集为 $\boldsymbol{S}=\{(y_i,\boldsymbol{x}_i),i=1,2...,M\}$，其中 y_i 为第 i 个样本的类别，$\boldsymbol{x}_i$ 为第 i 个样本的特征向量，M 为样本总数。当 N 个基分类器对数据集 $\boldsymbol{S}$ 依次进行 k 折交叉验证时，对于测试集中的每个样本 x_i，基分类器的预测结果为 $\boldsymbol{z}_{Ni}$，将每个基分类器的 k 次测试结果合并与原始数据标签 y_i 一起构成元分类器的输入向量即 $\boldsymbol{S}_{new}=\{(y_i,z_{1i},z_{2i},...,z_{Ni}),i=1,2,...,M\}$。元分类器通过学习新构成的数据特征 $\boldsymbol{S}_{new}$，输出最终判别属性。

对于 Stacking 集成学习而言，基分类器和元分类器的设计是关键之处。LR、KNN 因为理论成熟、简单高效等优点，在很多领域有着很好的应用效果[31-32]。RF、GBDT 分别是基于 Bagging 和 Boosting 的思想，RF 能够高度并行化训练，大大提高了计算效率，而 GBDT 通过构造弱分类器，使每个模型输出结果与残差之和尽可能的与预测值接近。从偏差和方差的角度分析，RF 主要降低误差的方差项，GBDT 既能降低偏差也能降低方差，这两个算法相互组合通过不同的机制能确保结果的有效性[33-34]。所以，本文采用了经典机器学习算法 LR、KNN、RF、GBDT 作为基分类器，元分类器使用泛化能力强的 GBDT 算法来纠正多个算法对训练集的偏置情况。

2.2 具体案例分析

在 Stacking 集成学习训练过程中如果直接用第一层模型的训练数据当做第二层元学习器的输入，可能会有过拟合的风险，因此需要对训练过程进行重新设计，为此针对金线莲品系识别提出了如图 3 所示的模型训练过程。主要步骤如下：

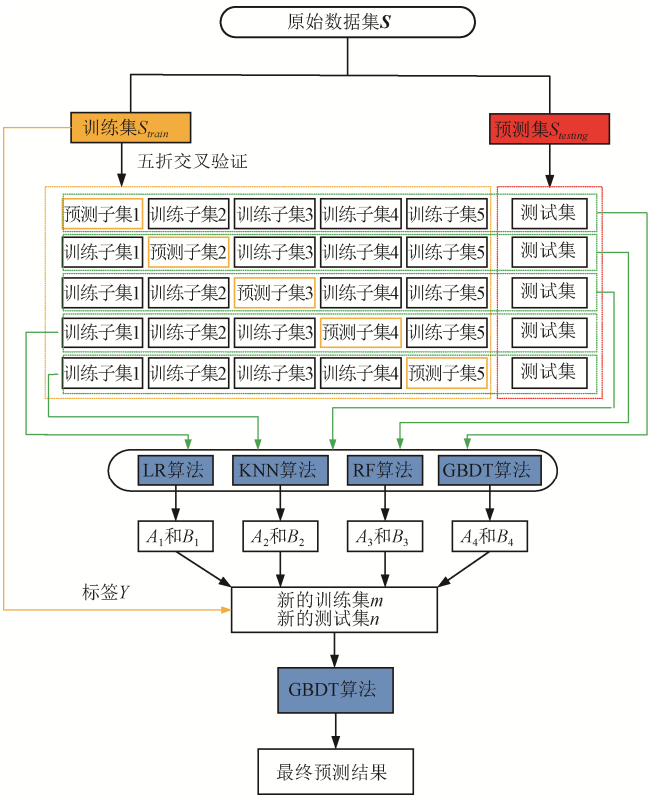


图 3 Stacking 框架下的金线莲品系识别

Fig.3 Strains identification of *Anoectochilus roxburghii* using the Stacking framework

步骤一：将金线莲样本数据集 $\boldsymbol{S}$ 按照 6∶4 的比例划分为训练集 S_{train}（190 个样本）和预测集 $S_{testing}$（127 个样本），将 S_{train} 按照五折交叉验证的方法随机均等划分为 5 个子集 $S_1, S_2, \cdots, S_5$，依次选取其中的一个子集 $S_i (i=1,2,\cdots,5)$ 作为预测子集，其他 $S_{-i} = S_{train} - S_i$ 作为训练子集。

步骤二：将 S_{-i} 作为基分类器 LR 的训练集，S_i 作为测试集，输出测试结果 α_i，同时对原始测试集 $S_{testing}$ 进行预测，输出预测结果 β_i。

步骤三：对步骤二循环 5 次得到 $\{\alpha_1, \alpha_2, \alpha_3, \alpha_4, \alpha_5\}$，将这 5 次的结果按照列合并得到和原始训练集 S_{train} 标签 Y 相同长度的列向量 $\boldsymbol{A}_1$，对预测样本值 $\{\beta_1, \beta_2, \beta_3, \beta_4, \beta_5\}$ 取平均值得到和 $S_{testing}$ 标签 Y 相同长度的列向量 $\boldsymbol{B}_1$。

步骤四：通过对另外 3 个基分类器 KNN、RF、GBDT 依次执行以上步骤得到由原始训练集 S_{train} 产生的 A_2, A_3, A_4 和原始测试集 $S_{testing}$ 产生的 B_2, B_3, B_4。

步骤五：将 A_1, A_2, A_3, A_4 和原始训练集 S_{train} 的标签 Y 合并得到的新样本数据 $\boldsymbol{m} = \{(A_1, A_2, A_3, A_4, Y)\}$ 作为元分类器 GBDT 的输入特征，$\boldsymbol{n} = \{(B_1, B_2, B_3, B_4)\}$ 作为元分类器的测试集来生成最终结果。

由于元分类器的训练集和测试集均未参与到各个基分类器的训练过程，所以有效防止了过拟合的发生，并且元分类器综合了各个基分类器的输出特征提升了分类准确率。

3 结果与讨论

3.1 试验平台与参数设置

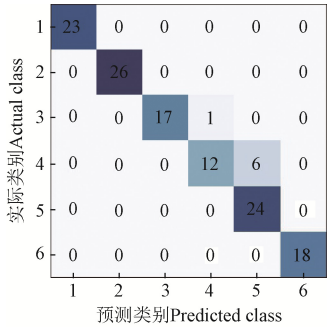
本次试验数据包括 6 个不同品系的 317 个金线莲图片，对样本进行子空间划分和特征提取形成 317 个 114 维的样本空间。试验在 Pycharm 环境下操作，所用的电脑系统配置为：Win10（64 位）操作系统，运行内存为 8 G，处理器为酷睿 i5-8250U，CPU 主频为 1.6 GHz，四核八线程。LR 算法选择 L_2 为正则项，分类方式采用 one-vs-rest 策略；KNN 算法最近邻个数 K 设置为 4；RF 算法叶子结点上的最小样本数量为 1，分割内部节点所最小样本数量为 2；决策树数目为 60，GBDT 算法的学习率为 0.01，弱学习器的最大迭代次数为 100，树的深度为 3。

3.2 鉴别结果分析

3.2.1 Stacking 集成学习模型鉴别结果分析

子区间分割之前，对叶片进行形状特征提取以此来验证金线莲叶片形态区分度小这一特点。本次试验选择 Hu 不变矩特征[35]和 6 个几何特征[36]（狭长度、矩形度、圆形度、周长长宽比、周长直径比、偏心率）一起作为 Stacking 模型的输入特征，试验结果表明叶片形状特征的分类正确率仅为 49.61%，故形状特征不能作为叶片有效识别特征，从而采用子区间分割技术来提高识别率。

子区间分割后，采用混淆矩阵来评估 Stacking 集成学习的性能，具体结果如图 4 所示。矩阵的每一行代表样本的实际类别，每一列代表预测类别结果，矩阵对角线上的数字代表每个品系被正确识别的样本数目。可以看出，台湾金线莲、红霞、一株圆叶、大圆叶这 4 个不同品系全部被正确识别，然而小圆叶品系中有 1 个样本被错误识别成尖叶，尖叶中有 6 个样本被误识别成一株圆叶，说明尖叶在纹理、颜色方面与人工自主培育品种一株圆叶有很大的相似性。



注：坐标轴数字代表 1.台湾金线莲 2.红霞 3.小圆叶 4.尖叶 5.一株圆叶 6.大圆叶
Note: Axis number representations 1. Taiwan 2. Hongxia 3. Xiaoyuanye 4. Jianye 5.Yizhu 6. Dayuanye

图 4 Stacking 模型的混淆矩阵

Fig.4 Confusion matrix of the Stacking model

此外，还使用了常见分类性能度量指标来评价分类模型，这些指标包括准确率 A（Accuracy）、精确率 P

（Precision）、召回率 R（Recall）、综合评价指标 F 值（F1-Score），公式如下：

$$A=\frac{\mathrm{TP+TN}}{\mathrm{TP+TN+FP+FN}} \tag{13}$$

$$P=\frac{\mathrm{TP}}{\mathrm{TP+FP}} \tag{14}$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP+FN}} \tag{15}$$

$$F=\frac{2PR}{P+R} \tag{16}$$

其中 TP 表示实际为正类的样本预测为正类的样本数，TN 表示实际为负类的样本预测为负类的样本数，FP 表示实际为负类的样本预测为正类的样本数，FN 表示实际为正类的样本预测为负类的样本数。分类器进行分类任务时，把要预测品系的实际类别数视为正样本数，其他所有品系之和为负样本数。如图 4 所示当对类别 3 进行预测分类时，正样本数为类别 3 的数目 18，负样本数为其他 5 个类别数之和 109，所以 TP =17，TN =103，FP =0，FN =1。类别 5 中，正样本数为 24，负样本数为 103，TP =24，TN =96，FP =6，FN =0。

根据公式（13），Stacking 模型下金线莲品系识别正确率 A（Accuracy）达到 94.49%，而 LR、KNN、RF、GBDT 的正确率分别为 89.13%、83.15%、87.56%、82.36%，Stacking 集成学习在一定程度上提升了分类准确性。

3.2.2　不同模型分类结果对比

为了充分验证构建的 Stacking 集成学习模型的有效性，根据式（14）～（16），针对每个不同品系将集成学习模型与单一分类模型分别在精确率 P、召回率 R 和综合评价指标 F 值方面进行对比分析，并归纳出分类器的整体识别能力。对比具体结果如表 1～4 所示。

表 1　不同分类模型精确率比较

Table 1 Precision comparisons of different classification models　%

类　别 Class	分类模型　Classification model				
	LR	KNN	RF	GBDT	Stacking
1	100.0	92.44	98.30	94.54	100.0
2	99.26	94.17	99.26	83.78	100.0
3	89.23	85.40	88.55	92.94	100.0
4	67.94	60.66	71.08	79.40	92.31
5	77.23	75.70	72.16	74.80	80.00
6	93.90	94.35	95.29	75.08	100.0

表 2　不同分类模型的召回率比较

Table 2　Recall comparisons of different classification models　%

类　别 Class	分类模型　Classification model				
	LR	KNN	RF	GBDT	Stacking
1	100.0	93.31	99.13	90.44	100.0
2	100.0	97.69	99.23	95.38	100.0
3	100.0	96.66	91.11	73.33	94.44
4	55.56	76.67	63.33	64.45	66.67
5	75.83	69.17	79.17	95.00	100.0
6	100.0	60.00	87.78	63.34	100.0

表 3　不同分类模型的综合评价指标比较

Table 3　F1-Score comparisons of different classification models　%

类　别 Class	分类模型　Classification model				
	LR	KNN	RF	GBDT	Stacking
1	100.0	93.09	98.70	92.40	100.0
2	99.62	95.86	99.23	88.82	100.0
3	94.28	90.63	89.54	81.92	97.14
4	69.94	67.65	66.38	70.73	77.42
5	76.40	72.23	75.02	83.44	88.89
6	96.82	72.75	68.52	68.52	100.0

表 4　不同分类模型整体平均识别能力比较

Table 4　Overall average recognition ability comparisons of different classification models　%

指标 Indexes	分类模型　Classification model				
	LR	KNN	RF	GBDT	Stacking
正确率 Accuracy	89.13	83.15	87.56	82.36	94.49
精确率 Precision	87.93	83.79	87.44	83.42	95.39
召回率 Recall	88.57	82.25	86.63	80.32	93.52
综合评价指标 F1-Score	89.51	82.04	82.90	80.97	93.91

从表 1～4 可以看出，Stacking 表现出良好的分类性能，在类别 1、类别 2、类别 6 品系识别中精确率 P、召回率 R、综合评价指标 F 值均达 100%。与单一分类模型相比，Stacking 集成模型在金线莲各个品系识别中均拥有最高的精确率，整体平均精确率达 95.39%，比单一分类模型中最高的 LR 模型高 7.46 个百分点。在召回率的比较中，Stacking 模型在类别 3 的召回率为 94.44%低于 LR 的 100%和 KNN 的 96.66%，在类别 4 上为 66.67%低于 KNN 模型的 76.67%，但在整体上达到 93.52%优于其中任何一个分类器。F 值作为综合评价指标表示分类器的整体性能，在每个品系识别中 Stacking 集成模型的 F 值均最高，且整体平均值达 93.91%，分别比 LR、KNN、RF、GBDT 这 4 个单一模型高出 4.40、11.87、11.01、12.94 个百分点，表现出该模型良好的综合性能。

以上结果表明，Stacking 集成模型综合了其他 4 种弱分类器分类性能，整体上拥有比单一分类器更好的表现性能，能够更加有效识别不同品系的金线莲。这是因为单一分类器在训练过程中往往可能陷入局部最优点，而局部最优对应的模型泛化性能不佳，Stacking 集成学习通过结合基分类器来有效减少陷入局部最优点的风险。然而，Stacking 集成学习在类别 3 与类别 4 的召回率表现上有所欠缺，这是因为 Stacking 模型受到基分类器的精度与多样性影响的原因，其基分类器之间的相关性大，在数据空间中获得的分类假设函数相似，无法在最大程度上体现不同基分类器的优势。

4　结　论

本文提出了多模型融合的 Stacking 集成学习方法，对 6 种不同品系的金线莲叶片进行分类识别。首先对原始图像进行灰度化、滤波降噪、阈值分割、形态学处理

等预处理，得到叶片二值图像，在二值图像的基础上提出叶片子区间分割方法，得到不同叶片相同部位的子区间目标图像。然后使用局部二进制模式（Local Binary Pattern，LBP）、灰度共生矩阵（Gray Level Co-occurrence Matrix，GLCM）、Gabor 滤波等方法提取图像纹理特征，使用颜色低阶矩提取颜色特征，这些特征相互融合作为金线莲的识别特征。为了充分学习数据特征，构建了以逻辑回归（Logistic Regression，LR）、K 最近邻（K Nearest Neighbor，KNN）、随机森林（Random Forest，RF）和梯度提升决策树（Gradient Boosting Decision Tree，GBDT）为基分类器的 Stacking 分类模型，以此来提升分类性能。通过试验研究比较，Stacking 模型在分类准确率、精确率、召回率和综合评价指标 F 值等指标上分别达 94.49%、95.39%、93.52%和 93.91%，优于 LR、KNN、RF 和 GBDT 模型。在今后的研究中，将进一步考虑 Stacking 配置选择方法，减少计算时间的情况下提高学习效率，同时将进一步扩充图像数据集，在更加复杂的背景下实现不同品系的植物识别。

[参 考 文 献]

[1] 邵清松，叶申怡，周爱存，等. 金线莲种苗繁育及栽培模式研究现状与展望[J]. 中国中药杂志，2016,41(2)：160-166.
Shao Qingsong, Ye Shenyi, Zhou Aicun, et al. Current researches and prospects of seedling propagation and cultivation modes of Jinxianlian[J]. China Journal of Chinese Materia Medica, 2016, 41(2): 160-166. (in Chinese with English abstract)

[2] Ye S, Shao Q, Zhang A. Anoectochilus roxburghii: A review of its phytochemistry, pharmacology, and clinical applications[J]. Journal of Ethnopharmacology, 2017, 209: 184-202.

[3] Tseng C C, Shang H F, Wang L F, et al. Antitumor and immunostimulating effects of Anoectochilus formosanus Hayata[J]. Phytomedicine, 2006, 13(5): 366-370.

[4] 陈莹，任丽，严桂杰，等. 不同来源金线莲的 HPLC 指纹图谱[J]. 沈阳药科大学学报，2019，36(9)：794-804.
Chen Ying, Ren Li, Yan Guijie, et al. Fingerprint analysis of Anoectochilus roxburghii from different sources by HPLC[J]. Journal of Shenyang Pharmaceutical University, 2019, 36(9): 794-804. (in Chinese with English abstract)

[5] Lv T, Teng R, Shao Q, et al. DNA barcodes for the identification of Anoectochilus roxburghii and its adulterants[J]. Planta, 2015, 242(5): 1167-1174.

[6] Tao O, Lin Z, Zhang X B, et al. Research on identification model of chinese herbal medicine by texture feature parameter of transverse section image[J]. World Science and Technology-Modernization of Traditional Chinese Medicine, 2014 (12): 2558-2562.

[7] 朱黎辉，李晓宁，张莹，等. 基于形状特征及纹理特征的中药材检索方法[J]. 计算机工程与设计，2014，35(11)：3903-3907.
Zhu Lihui, Li Xiaoning, Zhang Ying, et al. Image retrieval method for Chinese herbal medicine based on shape features and texture features[J]. Computer Engineering and Design, 2014, 35(11): 3903-3907. (in Chinese with English abstract)

[8] 宋彦，谢汉垒，宁井铭，等. 基于机器视觉形状特征参数的祁门红茶等级识别[J]. 农业工程学报，2018，34(23)：279-286.
Song Yan, Xie Hanlei, Ning Jingming, et al. Grading Keemun black tea based on shape feature parameters of machine vision[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2018, 34(23): 279-286. (in Chinese with English abstract)

[9] Turkoglu M, Hanbay D. Leaf-based plant species recognition based on improved local binary pattern and extreme learning machine[J]. Physica A: Statistical Mechanics and its Applications, 2019, 527: 121297.

[10] 张凯兵，章爱群，李春生. 基于 HSV 空间颜色直方图的油菜叶片缺素诊断[J]. 农业工程学报，2016，32(19)：179-187.
Zhang Kaibing, Zhang Aiqun, Li Chunsheng. Nutrient deficiency diagnosis method for rape leaves using color histogram on HSV space[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2016, 32(19): 179-187. (in Chinese with English abstract)

[11] 邓向武，齐龙，马旭，等. 基于多特征融合和深度置信网络的稻田苗期杂草识别[J]. 农业工程学报，2018，34(14)：165-172.
Deng Xiangwu, Qi Long, Ma Xu, et al. Recognition of weeds at seedling stage in paddy fields using multi-feature fusion and deep belief networks[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2018, 34(14): 165-172. (in Chinese with English abstract)

[12] Sharma P, Aggarwal A, Gupta A, et al. Leaf identification using HOG, KNN, and neural networks[C]//International Conference on Innovative Computing and Communications. Springer, Singapore, 2019: 83-91.

[13] 马娜，李艳文，徐苗. 基于改进 SVM 算法的植物叶片分类研究[J]. 山西农业大学学报：自然科学版，2018，38(11)：39-44.
Ma Na, Li Yanwen, Xu Miao. Plant leaf classification using improved SVM algorithm[J]. Journal of Shanxi Agricultural University: Natural Science Edition, 2018, 38(11): 39-44. (in Chinese with English abstract)

[14] 刘立波，程晓龙，戴建国，等. 基于逻辑回归算法的复杂背景棉田冠层图像自适应阈值分割[J]. 农业工程学报，2017，33(12)：201-208.
Liu Libo, Cheng Xiaolong, Dai Jianguo, et al. Adaptive threshold segmentation for cotton canopy image in complex background based on logistic regression algorithm[J]. Transactions of the Chinese Society of Agricultural

Engineering (Transactions of the CSAE), 2017, 33(12): 201-208. (in Chinese with English abstract)

[15] Dong X, Yu Z, Cao W, et al. A survey on ensemble learning[J]. Frontiers of Computer Science, 2020, 14(2): 241-258.

[16] Andiojaya A, Demirhan H. A bagging algorithm for the imputation of missing values in time series[J]. Expert Systems with Application, 2019, 129(9): 10-26.

[17] Wang B, Pineau J. Online bagging and boosting for imbalanced data streams[J]. IEEE Transactions on Knowledge and Data Engineering, 2016, 28(12): 3353-3366.

[18] Hui Y, Mei X, Jiang G, et al. Milling tool wear state recognition by vibration signal using a stacked generalization ensemble model[J]. Shock and Vibration, 2019(3): 1-16.

[19] Elayidom S, Idikkula S M, Alexander J. A hybrid stacking ensemble framwork for employment predicyion problems[J]. Advances in Computational Research, 2011, 3(1): 25-30

[20] Dinakar K, Weinstein E, Lieberman H, et al. Stacked generalization learning to analyze teenage distress[C]// 2014 8th International AAAI Conference on Weblogs and Social Media (ICWSM 2014), Ann Arbor, Michigan, USA, 2014: 81-90

[21] Haddad B M, Yang S, Karam L J, et al. Multifeature, sparse-based approach for defects detection and classification in semiconductor units[J]. IEEE Transactions on Automation Science and Engineering, 2016, 15(1): 1-15.

[22] 袁培森，杨承林，宋玉红，等. 基于 Stacking 集成学习的水稻表型组学实体分类研究[J]. 农业机械学报，2019, 50(11)：144-152.

Yuan Peisen, Yang Chenglin, Song Yuhong, et al. Classification of rice phenomics entities based on stacking ensemble learning[J]. Transactions of the Chinese Society for Agricultural Machinery, 2019, 50(11): 144-152. (in Chinese with English abstract)

[23] 王丽爱，周旭东，朱新开，等. 基于 HJ-CCD 数据和随机森林算法的小麦叶面积指数反演[J]. 农业工程学报，2016，32(3)：149-154.

Wang Liai, Zhou Xudong, Zhu Xinkai, et al. Inverting wheat leaf area index based on HJ-CCD remote sensing data and random forest algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2016, 32(3): 149-154. (in Chinese with English abstract)

[24] 张会清，牛铮. 基于线性判别分析和梯度提升决策树的 WLAN 室内定位算法[J]. 仪器仪表学报，2018，39(12)：136-143.

Zhang Huiqing, Niu Zheng. WLAN indoor positioning algorithm based on linear discriminant analysis and gradient boosting decision tree[J]. Chinese Journal of Scientific Instrument. 2018, 39(12): 136-143. (in Chinese with English abstract)

[25] 王见，周勤，尹爱军. 改进 Otsu 算法与 ELM 融合的自然场景棉桃自适应分割方法[J]. 农业工程学报，2018，34(1)：173-180.

Wang Jian, Zhou Qin, Yin Aijun. Self-adaptive segmentation method of cotton in natural scene by combining improved Otsu with ELM algorithm[J]. Transactions of the Chinese Society of Agricultural Engineering (Transactions of the CSAE), 2018, 34(1): 173-180. (in Chinese with English abstract)

[26] Naresh Y G, Nagendraswamy H S. Classification of medicinal plants: An approach using modified LBP with symbolic representation[J]. Neurocomputing, 2016, 173: 1789-1797.

[27] Wu Q, Gan Y, Lin B, et al. An active contour model based on fused texture features for image segmentation[J]. Neurocomputing, 2015, 151: 1133-1141.

[28] VijayaLakshmi B, Mohan V. Kernel-based PSO and FRVM: An automatic plant leaf type detection using texture, shape, and color features[J]. Computers and Electronics in Agriculture, 2016, 125: 99-112.

[29] Yang H, Yin J, Jiang M. Perceptual image hashing using latent low-rank representation and uniform LBP[J]. Applied Sciences, 2018, 8(2): 317.

[30] Das S, Rudrapal D. Analysis of color moment as a low level feature in improvement of content based image retrieval[C]// Proceedings of the Fourth International Conference on Signal and Image Processing 2012 (ICSIP 2012). Springer, India, 2013: 387-397.

[31] Rymarczyk T, Kozowski E, Kosowski G, et al. Logistic regression for machine learning in process tomography[J]. Sensors, 2019, 19(15): 3400.

[32] Lee T R, Wood W T, Phrampus B J. A machine learning (KNN) approach to predicting global seafloor total organic carbon[J]. Global Biogeochemical Cycles, 2019, 33(1): 37-46

[33] Gui L, Xia Y, Li H, et al. Prediction of NOX emission from coal - fired boiler based on RF - GBDT[C]// 2017 6th International Conference on Energy and Environmental Protection (ICEEP 2017). 2017: 344-350.

[34] 徐兵，刘潇，汪子扬，等. 采用梯度提升决策树的车辆换道融合决策模型[J]. 浙江大学学报：工学版，2019，53(6)：158-168.

Xu Bing, Liu Xiao, Wang Ziyang, et al. Fusion decision model for vehicle lane change with gradient boosting decision tree[J]. Journal of Zhejiang University: Engineering Science, 2019, 53(6): 158-168. (in Chinese with English abstract)

[35] Wang X F, Huang D S, Du J X, et al. Classification of plant leaf images with complicated background[J]. Applied Mathematics & Computation, 2008, 205(2): 916-926.

[36] Ahmed F, Almamun H A, Bari A S, et al. Classification of crops and weeds from digital images: A support vector machine approach[J]. Crop Protection, 2012, 40: 98-108.

Strains classification of *Anoectochilus roxburghii* using multi-feature extraction and Stacking ensemble learning

Xie Wenyong[1,2], Chai Qinqin[1,2※], Gan Yonghui[3], Chen Shudi[1,2], Zhang Xun[4], Wang Wu[1,2]

(1. *College of Electrical Engineering and Automation*, *Fuzhou University*, *Fuzhou* 350108, *China*; 2. *Ministry of Education Key Laboratory of Medical Instrument and Pharmaceutical Technology*, *Fuzhou University*, *Fuzhou* 350108, *China*; 3. *School of Food Engineering*, *Zhangzhou Institute of Technology*, *Zhangzhou* 363000, *China*; 4. *College of Pharmacy*, *Fujian University of Traditional Chinese Medicine*, *Fuzhou* 350122, *China*)

Abstract: *Anoectochilus roxburghii* (A.*roxburghii*) is a rare medicinal herb that mainly distributed in China. It is necessary to identify strains of A.*roxburghii* for the guidance of clinical medication, due to different strains distinctly vary in medicinal values. However, similar leaf morphology has made difficult to discern different strains directly by naked eyes. In this study, a sub-interval segmentation method was proposed to identify the different strains of A.*roxburghii*, based on leaf identification methods. Firstly, 6 strains of A.*roxburghii* were selected, including Taiwan, Hongxia, Xiaoyuanye, Jianye, Yizhu, Dayuanye. A total of 317 images with the resolution of 800×800 pixels were taken, while two filtering methods were used to remove noise. The maximum inner variance algorithm was used for automatic threshold segmentation, in order to obtain the binary image. In the binary image, the leaf contour was drawn, and the mass center of the leaf was calculated. The square area with 150 pixels centered on the mass center was selected as the sub-interval of the leaf, to obtain the target image with the same position and size. Secondly, a combination of texture and color features was applied for the target image, in which texture features were derived by local binary patterns (LBP), gray level co-occurrence matrix (GLCM) and gabor filters, whereas, the color feature was composed of the first, second and third moments. After that, 114 merged features were obtained. Thirdly, the stacking ensemble learning was proposed to improve the accuracy of traditional single classifier. The stacking framework consisted of a base classifier, and a meta-classifier. Logistic regression (LR), K nearest neighbor (KNN), random forest (RF), and gradient boosting decision tree (GBDT) were used as the base classifiers, whereas, GBDT was used as the meta-classifier for stacking. Finally, the cross-validation method different from conventional model was used to divide the data set. The original data was normalized and randomly segmented, where 60% for training and 40% for testing. The training data set was randomly divided into 5 training subsets, and then testing subset for training each base classifier. The prediction results of base classifiers were used as the input vectors of the GBDT. The final prediction result was output by GBDT. The experiment results showed that the average recognition accuracy of the stacking reached 94.49%, while that of LR, KNN, RF and GBDT was 89.13%, 83.15%, 87.56%, 82.36%, respectively. Moreover, the Precision, Recall, and *F*1-Score of the stacking model for the identification of Taiwan, Hongxia, and Dayuanye were all 100%. The Recall performance of stacking model was better than any of the single classifiers for identification of the Xiaoyuanye, just slightly worse than that of the LR and KNN models. The *F*1-Score of stacking model reached the maximum in each strain identification, showing the excellent overall performance of the model. Therefore, the proposed method can significantly improve the classification performances of A.*roxburghii* with different strains. The findings can provide a promising application method to recognize leaves of different plants using shape features. A further research is still necessary to select proper configuration, in order to improve learning efficiency of stacking model.

Keywords: computer vision; models; *Anoectochilus roxburghii*; sub-interval segmentation; feature extraction; Stacking ensemble learning; plant leaf